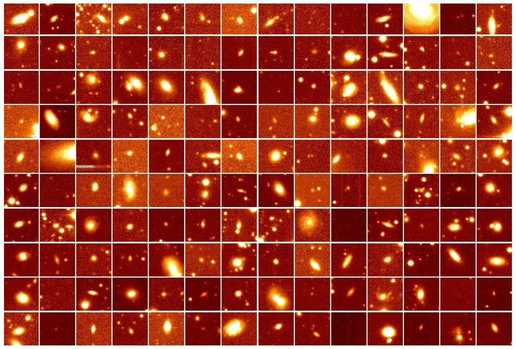
Hadoop for Scientific Workloads

Lavanya Ramakrishnan
Shane Canon
Shreyas Cholia
Keith Jackson
John Shalf
Lawrence Berkeley National Lab

YAHOO! PRESENTS
hadoop SUMMIT 2010

Example Scientific Applications

- Integrated Microbial Genomes (IMG)
  - analysis of microbial community metagenomes in the integrated context of all public reference isolate microbial genomes
- Supernova Factory
  - tools to measure expansion of universe and energy
  - task parallel workflow, large data volume



- MODerate-resolution Imaging Spectroradiometer (MODIS)
  - two MODIS satellites near polar orbits
  - ~ 35 science data products including atmospheric and land products
  - products are in different projection, resolutions (spatial and temporal), different times

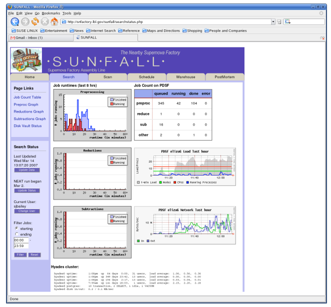
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

computational research division

YAHOO! PRESENTS hadoop SUMMIT 2010

Supporting Science at LBL

Scientists



User interfaces, grid middleware, workflow tools, data management, etc

HPC and IT resources

- Unlimited need for compute cycles and data storage
- Tools and middleware to access resources

Does cloud computing

- ➢make it easier or better to do what we do?
- ➢help us do things differently than before?
- ➢help us include other users?

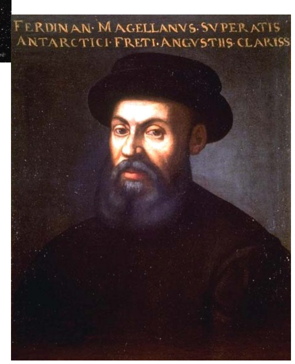
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

CRD computational research division

YAHOO! PRESENTS hadoop SUMMIT 2010

Magellan – Exploring Cloud Computing

- Test-bed to explore Cloud Computing for Science
- National Energy Research Scientific Computing Center (NERSC)
- Argonne Leadership Computing Facility (ALCF)
- Funded by DOE under the American Recovery and Reinvestment Act (ARRA)



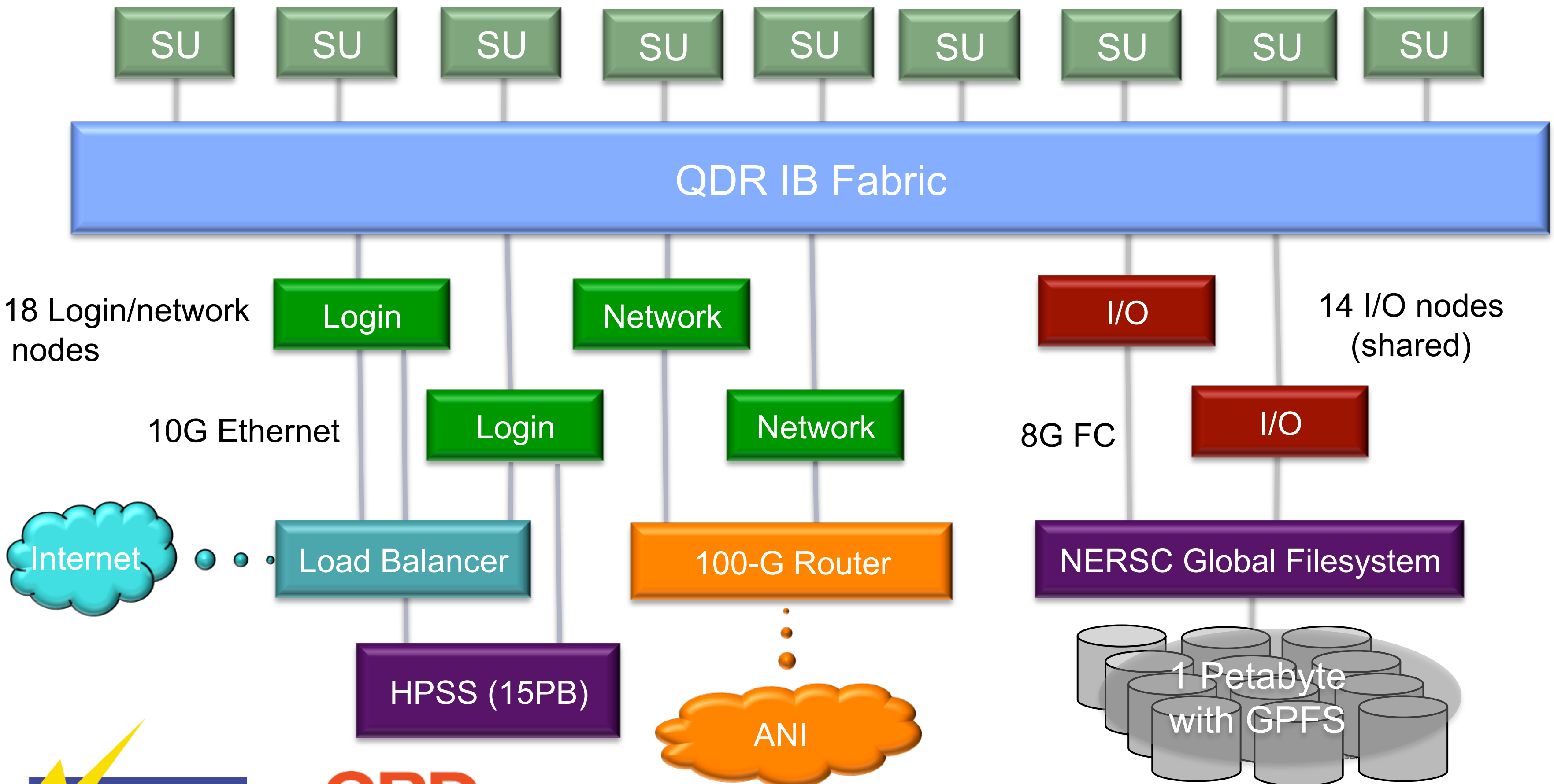
NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

CRD
computational research division

hadoop SUMMIT 2010

Magellan Cloud at NERSC

720 nodes, 5760 cores in 9 Scalable Units (SUs) ➔ 61.9 Teraflops
SU = IBM iDataplex rack with 640 Intel Nehalem cores



NERSC NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

CRD computational research division

hadoop SUMMIT 2010

Magellan Research Agenda

- What are the unique needs and features of a science cloud?
- What applications can efficiently run on a cloud?
- Are cloud computing programming models such as Hadoop effective for scientific applications?
- Can scientific applications use a data-as-a-service or software-as-a-service model?
- Is it practical to deploy a single logical cloud across multiple DOE sites?
- What are the security implications of user-controlled cloud images?
- What is the cost and energy efficiency of clouds?

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

computational research division

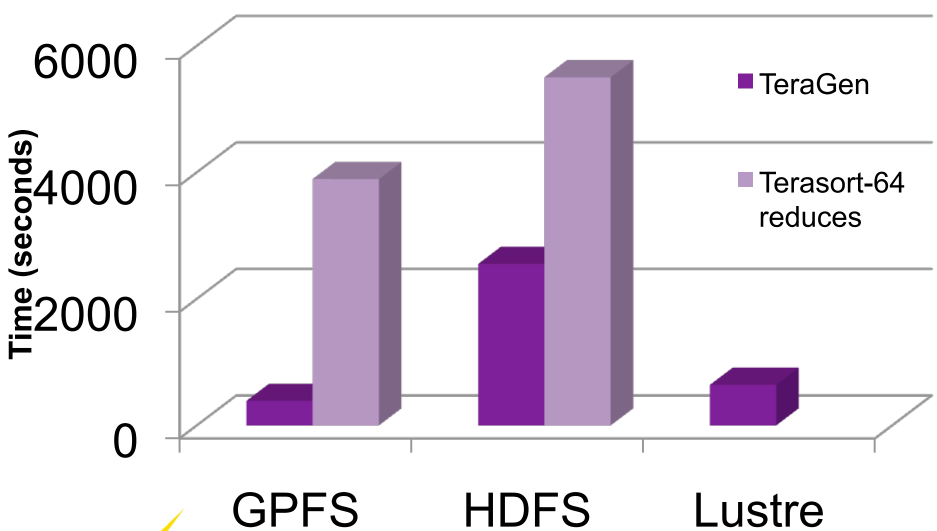
YAHOO! PRESENTS hadoop SUMMIT 2010

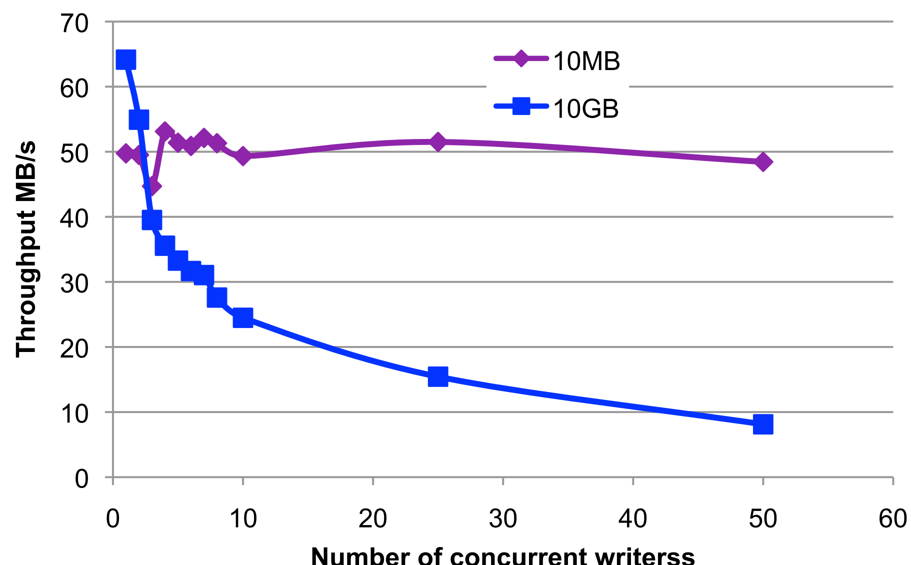
Hadoop for Science

- Classes of applications
  - tightly coupled MPI application, loosely couple data intensive science
  - use batch queue systems in supercomputing centers, local clusters and desktop
- Advantages of Hadoop
  - transparent data replication, data locality aware scheduling
  - fault tolerance capabilities
- Mode of operation
  - use streaming to launch a script that calls executable
  - HDFS for input, need shared file system for binary and database
  - input format
    - handle multi-line inputs (BLAST sequences), binary data (High Energy Physics)

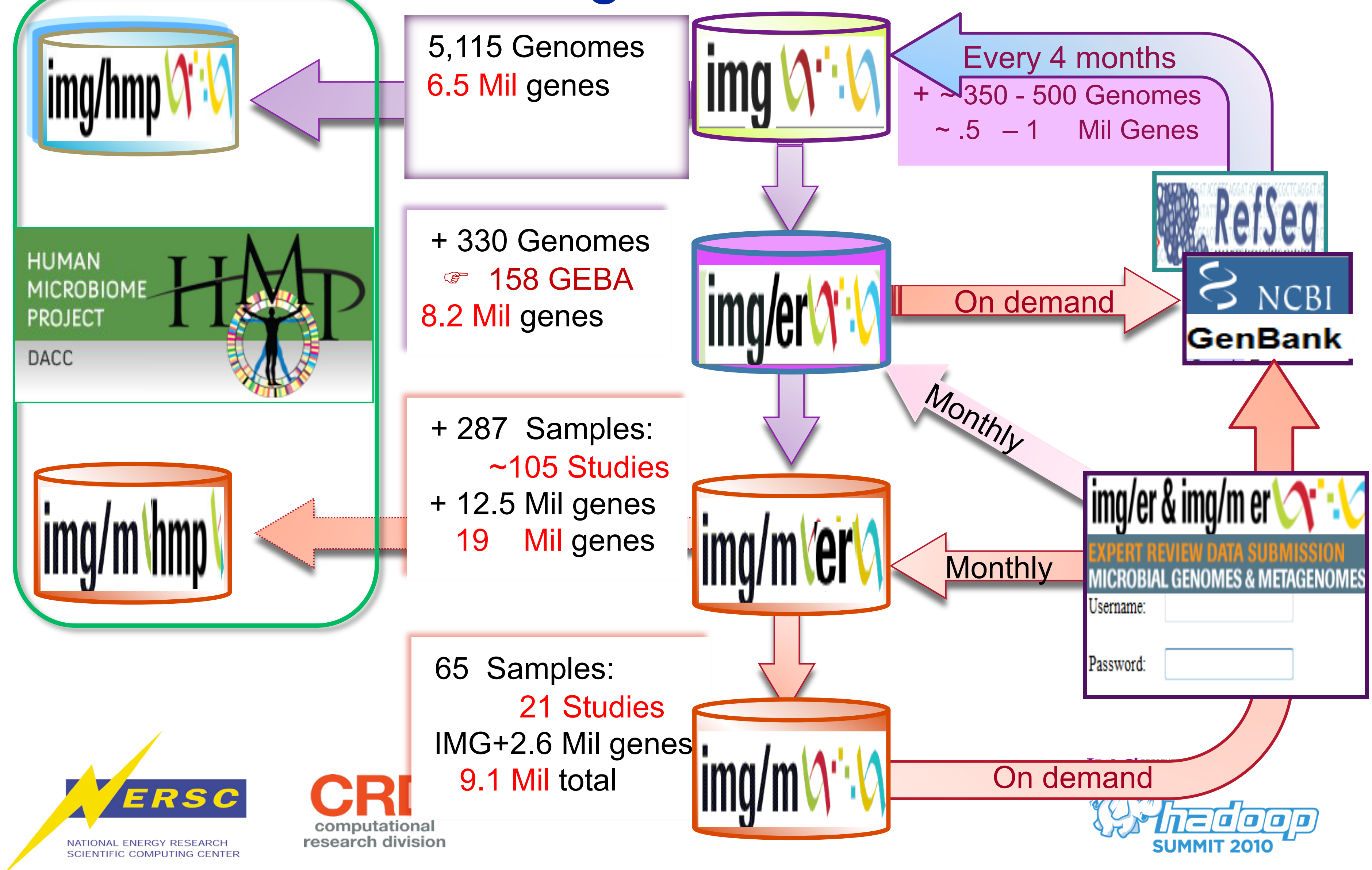
Hadoop Benchmarking: Early Results

- Compare traditional parallel file systems to HDFS
  - 40 node Hadoop cluster where each node contains two Intel Nehalem quad-core processors
  - TeraGen and Terasort to compare file system performance
    - 32 maps for TeraGen and 64 reduces for Terasort over a terabyte of data
  - TestDFSIO to understand concurrency





IMG Systems: Genome & Metagenome Data Flow



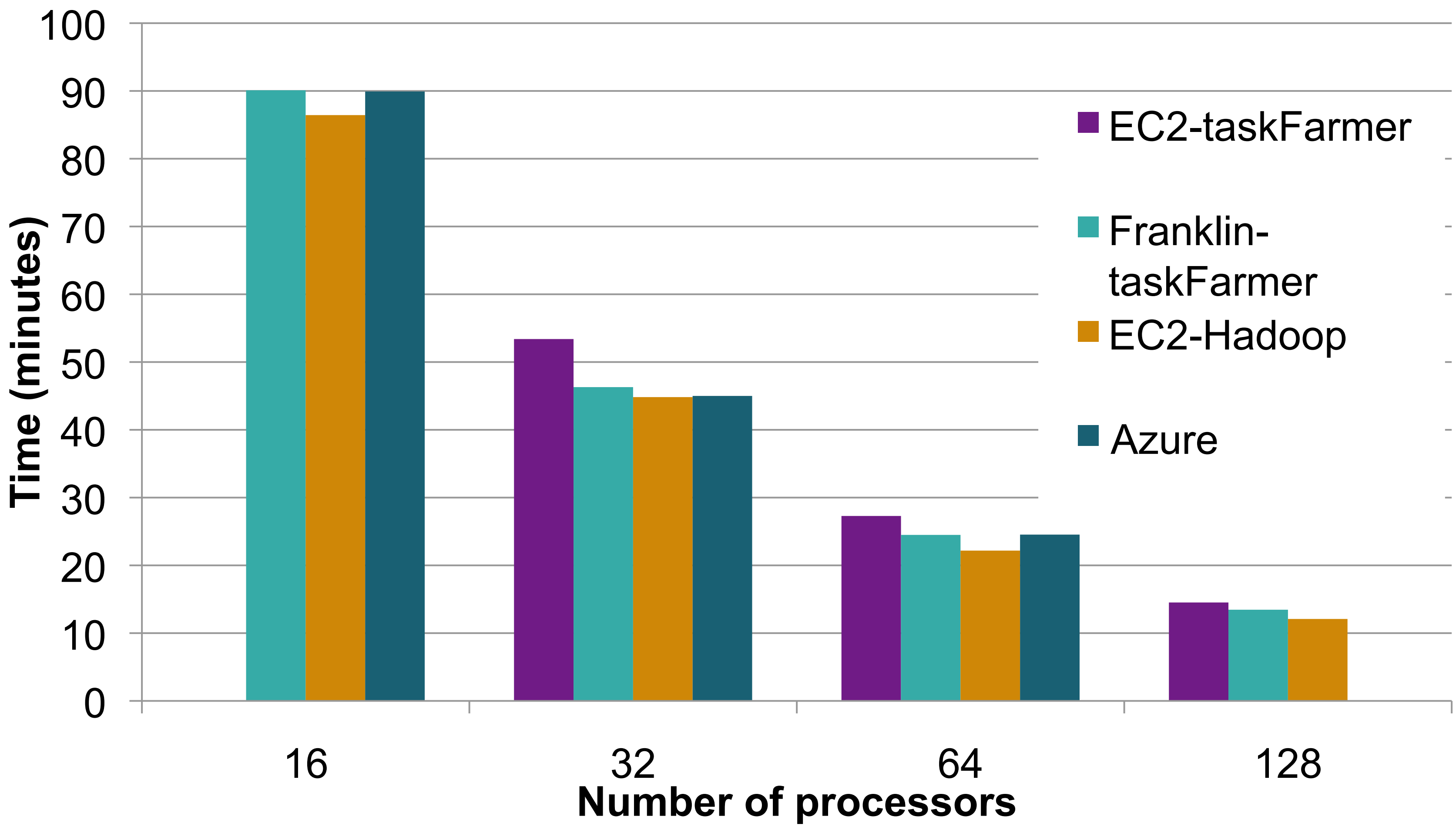
BLAST on Hadoop

- NCBI BLAST (2.2.22)
  - reference IMG genomes- of 6.5 mil genes (~3Gb in size)
  - full input set 12.5 mil metagenome genes against reference
- BLAST Hadoop
  - uses streaming to manage input data sequences
  - binary and databases on a shared file system
- BLAST Task Farming Implementation
  - server reads inputs and manages the tasks
  - client runs blast, copies database to local disk or ramdisk once on startup, pushes back results
  - advantages: fault-resilient and allows incremental expansion as resources come available

Hardware Platforms

- Franklin: Traditional HPC System
  - 40k core, 360TFLOP Cray XT4 system at NERSC, Lustre parallel filesystem
- Amazon EC2: Commercial “Infrastructure as a Service” Cloud
  - Configure and boot customized virtual machines in Cloud
- Yahoo M45: Shared Research “Platform as a Service” Cloud
  - 400 nodes, 8 cores per node, Intel Xeon E5320, 6GB per compute node, 910.95TB
  - Hadoop/MapReduce service: HDFS and shared file system
- Windows Azure BLAST “Software as a Service”

BLAST Performance



NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

computational research division

YAHOO! PRESENTS hadoop SUMMIT 2010

BLAST on Yahoo! M45 Hadoop

- Initial config – Hadoop memory ulimit issues,
  - Hadoop memory limits increased to accommodate high memory tasks
  - 1 map per node for high memory tasks to reduce contention
  - thrashing when DB does not fit in memory
- NFS shared file system for common DB
  - move DB to local nodes (copy to local /tmp).
  - initial copy takes 2 hours, but now BLAST job completes in < 10 minutes
  - performance is equivalent to other cloud environments.
  - future: Experiment with Distributed Cache
- Time to solution varies - no guarantee of simultaneous availability of resources

Strong user group and sysadmin support was key in working through this.

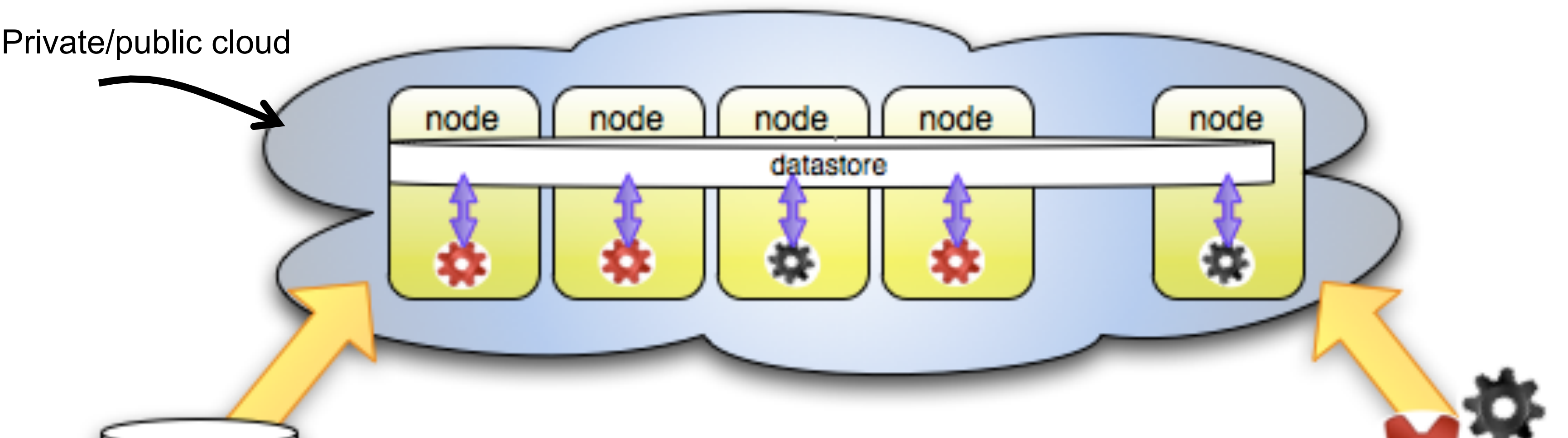
HBase for Metagenomics

- Output of “all vs. all” pairwise gene sequence comparisons
  - › currently data stored in compressed files
    - modifying individual entries is challenging
    - queries are hard
  - › duplication of data to ease presentation by different UI components
- Evaluating changing to Hbase
  - › easily update individual rows and simple queries
  - › query and update performance exceeds requirements
- Challenge: Bulk loads of approximately 30 billion rows
  - › trying multiple techniques for bulk loading
  - › best practices are not well documented

Magellan Application: De-novo assembly

Memory requirements: ~500 GB (de Bruijn graph)

CPU hours (single assembly): velveth: ~23h,velvetg: ~21h



- Move data from disk to clustered memory
- Move analysis pipeline from single-node to parallel map/reduce jobs

==

efficient horizontal scalability

(more data -> add more nodes)

Source: Karan Bhatia

Summary

- Deployment Challenges
  - all jobs run as user “hadoop” affecting file permissions
  - less control on how many nodes are used - affects allocation policies
  - file system performance for large file sizes
- Programming Challenges: No turn-key solution
  - using existing code bases, managing input formats and data
- Performance
  - BLAST over Hadoop: performance is comparable to existing systems
  - existing parallel file systems can be used through Hadoop On Demand
- Additional benchmarking, tuning needed
- Plug-ins for Science

Acknowledgements

This work was funded in part by the Advanced Scientific Computing Research (ASCR) in the DOE Office of Science under contract number DE-C02-05CH11231.

CITRIS/UC, Yahoo M45!, Greg Bell, Victor Markowitz, Rollin Thomas

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Questions?

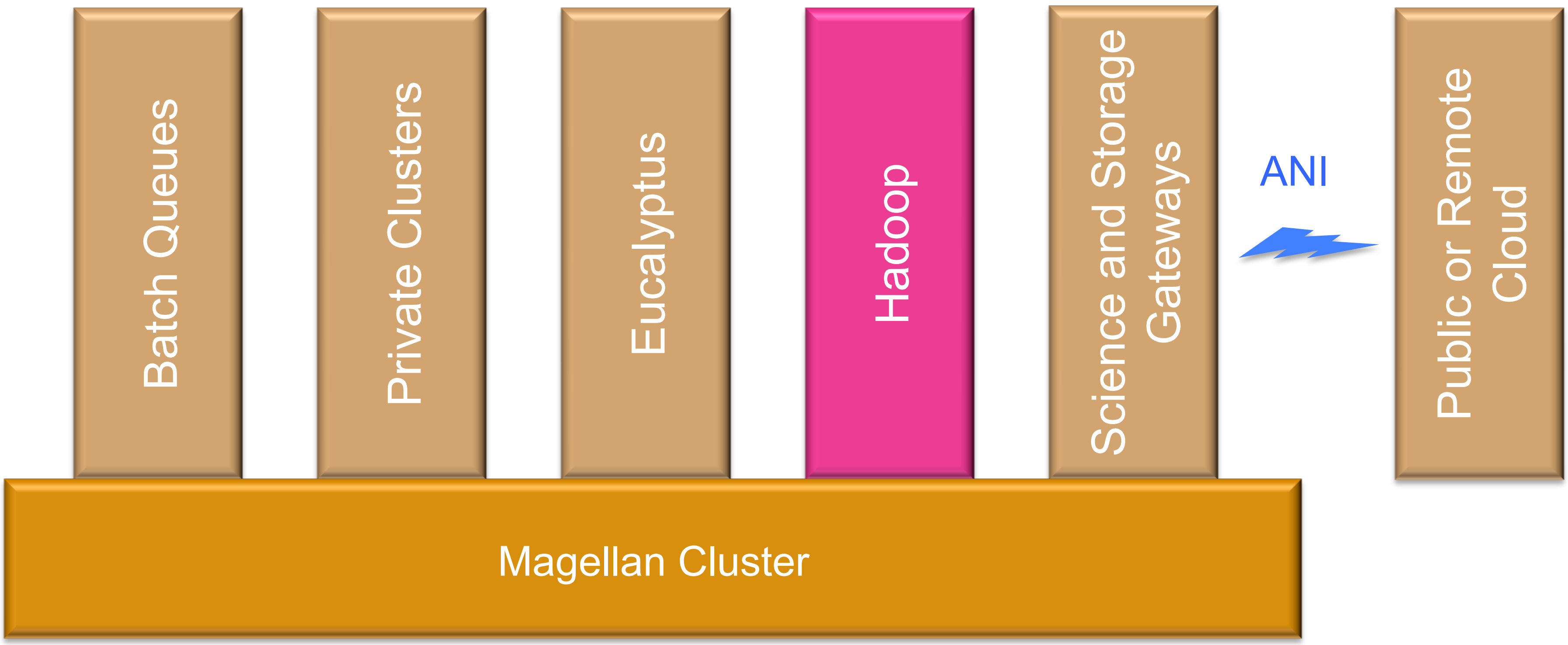
LRamakrishnan@lbl.gov

YAHOO! PRESENTS hadoop SUMMIT 2010

Cloud Usage Model

- On-demand access to computing and cost associativity
- Customized and controlled environments
  - e.g., Supernova Factory codes have sensitivity to OS/compiler versions
- Overflow capacity to supplement existing systems
  - e.g., Berkeley Water Center has analysis that far exceeds capacity of desktops
- Parallel programming models for data intensive science
  - e.g., BLAST parametric runs

NERSC Magellan Software Strategy



- Runtime provisioning of software images via Moab and xCat
- Explore a variety of usage models
- Choice of local or remote cloud